Individual internet usage and the availability of online content of local interest: a multilevel approach

Tranos, Emmanouil Stich, Christoph

July 31, 2019

Abstract

This paper illustrates whether the availability of online content of local interest affects the likelihood of individuals to connect to the internet and spend more time online. While the literature demonstrates a number of factors which *push* or enable individuals to spend more time online, we know little about the conditions that *pull* or attract individuals online. Although we know that individuals use the internet to access information, we do not know whether such attraction forces are relevant at the local scale too. Gaining a better understanding of how such mechanisms work at the local scale can assist our efforts to bridge digital divides, which tend to be geographically clustered. To explore this we utilise innovative data, which contain all the archived webpages under the UK top level domain name (.uk) and we calculate the volume of internet content of local interest at the neighbourhood level using the geolocation information included in the text of these webpages. Specifically, we calculate the radius of gyration for every archived website using the different postcodes included in the archived webpages and then we create an aggregated measure at the neighbourhood level discounting websites that have less of a local focus. We merge this measure of Local Internet Content (LIC) with a large population survey, which contains information about the frequency of internet usage in the UK and estimate the effect of LIC on the likelihood of an individual being a frequent internet user. Multilevel models are employed to utilise both individual and geographical level characteristics. Our results indicate that even after controlling for the individual and geographical characteristics, which according to previous studies affect internet usage, the availability of internet content of local interest still attracts individuals online.

Keywords: internet usage; internet archive; multilevel modelling

1 Introduction

Researchers have spent substantial effort in order to understand the mechanisms which influence the internet adoption and usage by individuals at different scales and levels of aggregation. Such studies, which come from various research fields including human geography, economics and media studies and are reviewed in the next section, have highlighted a number of driving factors related both to the socio-demographic characteristics of the individuals as well as the attributes of their location (e.g. urban vs. rural).

What has escaped these explorations is the opportunities that *pull* individuals online. Using an analogy from gravitational models, if, for example, high income and residency in urban locations act as *push* factors, which enable or influence individuals to connect to the internet and spend time online, researchers have spent less effort in understanding what attracts individuals to the internet and, more specifically, whether such attractions or, in other words, online opportunities have a spatial signature. The capacity of the internet to provide such online opportunities for accessing knowledge and information relevant at the *global* scale is well established and is reflected in its global reach (3 billion users) and capacity to support large scale one-to-many and many-to-many interactions (Graham et al., 2015b). However, we know little about whether opportunities for accessing knowledge and information relevant at the *local* scale can also attract individuals to connect to the internet and spend more time online. Gaining a better understanding of how these mechanisms work at the local scale can assist efforts to bridge digital divides, which tend to be geographically clustered (Hindman, 2000; Graham, 2011). Comprehending the relationship between internet usage and local online opportunities can also assist in supporting national economies to fully develop their digital potential (Bughin et al., 2016). For instance, gaining such an understanding can help firms to adjust their e-retail strategies in terms of pricing, advertising, and ranging products (Agarwal et al., 2009). From a broader perspective, by understanding the availability of such online opportunities at a local scale we are improving our capacity to depict digital augmentations of places, which nowadays tend to matter as much as much as their material counterparts (Graham et al., 2015b). As relevant research has illustrated, such digital augmentations of places can directly affect our economic, social and political interactions with material places (Kitchin and Dodge, 2011; Graham et al., 2015a).

This paper addresses this gap by utilising a novel source of data of all the archived webpages under the .uk domain by the Internet Archive, the most complete archive of webpages in the world (Holzmann et al., 2016; Ainsworth et al., 2011). To geolocate these data, the text from the archived

webpages has been scanned for the inclusion of a valid UK postcode and, therefore, we are able to build a measure of the volume of web content of local interest and test whether the availability of such web content acts as a pull factor for individuals to connect to the internet. To answer this research question, we spatially match this measure of Local Internet Content (LIC) with the responses of a large population survey for the UK. We then estimate multilevel models (MLM) to explain whether the likelihood of an individual to connect to the internet and spend time online is affected by individual and geographical factors. Going beyond the current state of the art, the geographical variables used here include a measure of the volume of LIC linked to the location of the survey participants.

Our research question is motivated by the nature of the web, which is the most notable element of the internet, as a techno-social system for human cognition, communication and co-operation, with cognition being the prerequisite and the precondition for the other two attributes (Fuchs et al., 2010). Cognition is the key attribute of Web 1.0 applications, namely static hypertext, but it is also present in Web 2.0 (e.g. Social Network Sites) and Web 3.0 applications (e.g. platforms such as GitHub and Stack Overflow), which also support communication and co-operation among internet users respectively. Hence, the availability of web content of local interest reflects a pool of information and knowledge about local opportunities which might be relevant to individuals residing in near proximity. Such opportunities can include anything from consumption opportunities of services and products from local businesses to the availability of services by local authorities and volunteer organisations. The literature suggests that internet content, among other socio-economic factors, has the capacity to stimulate internet adoption, at least at a national level (Viard and Economides, 2014). Moreover, Bekkerman and Gilpin (2013) suggest that increased internet access is associated with increased demand for information as reflected in library visits in the US. Interestingly, this pattern is stronger in metropolitan areas demonstrating how internet usage complements agglomeration forces. Nevertheless, the literature has not yet tested whether the availability of local internet content can pull individuals to engage more with the internet, a gap that this paper fulfils.

The structure of the paper goes as follows. The next section provides a review of the literature which explores the factors which drive individual engagement with the internet. Then, we introduce the novel data we use in this paper and Section 4 describes the methods employed. The results of the analysis are presented and discussed in Section 5. The paper ends with a conclusions section.

2 Literature Review

Inspired by the fascination of urban and economic geographers with telecommunications (e.g. Gottmann, 1977), human geography research spent a lot of effort in understanding the interrelation between internet and space. Most of the geographical research which studied the spatiality of the internet approached it from the supply side. Focusing on the highest tier of internet infrastructure Wheeler and O'Kelly (1999) analysed the topology of the internet's hardware and the derived city connectivities, Malecki (2002) focused on how urban hierarchies shaped the spatiality of internet's infrastructure and, more recently, Tranos (2013) analysed the geography and the spatial economic effects of the internet's main infrastructural networks in Europe. The tendency of this infrastructure to be concentrated in large metropolitan areas mirroring established globalisation patterns, but also challenging them in some cases, was the key finding from this strand of research.

The urban character of the internet is also reflected in the end-user broadband connectivity. US-focused research illustrated a core-periphery pattern in broadband provision (Grubesic, 2008), which is more complex than the traditional urban-rural or rich-poor spatial dichotomies (Mack and Grubesic, 2009). Similar patterns are also evident in the UK. Riddlesden and Singleton (2014) utilised broadband speed micro-data to explore the effect of urbanisation and population density on Internet broadband speed. Moreover, Oughton et al. (2015) confirmed that the previous findings regarding the internet infrastructure pull factors can also explain the spatial pattern of the internet broadband speed at the neighbourhood scale in the UK. There is also evidence that the adoption of internet and new related technologies follows an *innovation-diffusion* process with more urban, affluent, and younger areas acting as centres of innovation (Farag et al., 2007).

The above studies, as well as others not cited here, enabled us to gain a good understanding of the spatiality of the internet infrastructure, something which can be attributed, to a certain extent, to data availability for such infrastructure. Although such supply side measures can explain part of the variation of individual internet uptake and usage, our understanding of the above, especially at sub-national scales, is still far from comprehensive (Blank et al., 2018) also because of limited data availability regarding individual online behaviour. The literature suggests some key demographic and socio-economic variables, which shape internet usage, its frequency and its different typologies: age, gender, education and income (Blank et al., 2018; Van Deursen and Van Dijk, 2014). For example, the likelihood and the frequency of using the internet decreases with age and female adults tend to use the internet more often as a communication tool (Blank and Groselj,

2014; Blank et al., 2018; Zillien and Hargittai, 2009). However, men tend to be more active online (Blank and Groselj, 2014; Calenda and Meijer, 2009; Di Gennaro and Dutton, 2006; Hoffmann et al., 2015). Moreover, both income and education, which tend to be correlated, increase the likelihood of internet usage, but the literature suggests that internet users of higher socio-economic classes will engage more often in *capital-enhancing* internet activities, while users from lower socio-economic classes will use the internet in less productive ways (Zillien and Hargittai, 2009).

Empirical literature focusing on this issue has also highlighted the urban-rural divide as the main geographical reasoning behind differences in internet usage. Nevertheless, this divide can also be underpinned by other demographic and socio-economic spatial patterns, which may mirror urban-rural patterns, namely age, education, income, as well as internet connectivity (Hindman, 2000; Forman et al., 2018; Blank et al., 2018). For instance, Mills and Whitacre (2003) indicated that on the one hand, two thirds of the metropolitan divide in internet household adoption in the US can be attributed to socio-economic and demographic differences among households. On the other hand, a portion of the remaining one third of the metropolitan divide, which reflects place-based differences, may be associated with internet connectivity differences and digital infrastructural. It needs to be highlighted though that Mills and Whitacre (2003) utilised data from 2001, a period when internet technologies were very different than the ones present today (e.g. dial-up modems) and, therefore, their results might not be directly transferable to current conditions. In a different study, Blank et al. (2018) used micro-simulation to illustrate the importance of demographic characteristics as the driving force of digital divides. Going a step further they claimed that regional characteristics, expressed as regional fixed effects, are not statistically significant in explaining the proportion of internet users. However, regional dummies can be non-significant predictors of internet usage, but this does not mean that there is no unobserved intra-regional variation of internet usage that the regional fixed effects cannot account for. For instance, these models did not control for the quality of internet service at the local level, something which cannot be controlled with regional dummies as internet connectivity varies substantially within regions. Such variation has been explored in some very granular studies on the quality of the internet infrastructure in the UK, which were discussed above (Riddlesden and Singleton, 2014; Oughton et al., 2015). Interestingly, the literature suggests that locational characteristics such as the supply of internet service may affect internet adoption (Forman et al., 2018).

In accordance to the above literature, a number of studies created geodemographic classifications of how individuals engage with the internet. Long-

ley et al. (2008) developed a UK-wide household classification of the level of awareness of digital technologies and their perceived impact on capital formation and quality of life. Moreover, Longley and Singleton (2009) used public consultation in order to further improve the previous classification. Recent attempts utilised data from extensive surveys to produce more detailed and accurate geodemographic classifications (Riddlesden, 2014; Singleton et al., 2015).

Despite the different approaches, all studies cited above focused on exposing the demographic, socio-economic and geographic factors behind internet usage and behaviour. A limited number of studies aimed to understand how spillovers can affect internet usage and online behaviour (Forman et al., 2018). For instance, Agarwal et al. (2009) exposed how *peer effects* driven by the geographical proximity of individuals and their collocation within the same metropolitan areas can influence internet use. In addition, Sinai and Waldfogel (2004) revealed a bidirectional phenomenon. On the one hand, larger markets tend to have more online content of local interest, something which can attract more individuals online. This finding indicates a complementary relation between internet and urban agglomeration. On the other hand, holding local online content constant, individuals are less likely to use the internet in large areas, a finding which indicates a substitutional relation between internet and cities. The role of local content as a pull factor for individuals to use the internet was highlighted in a study that took place soon after the commercialisation of the internet. Kraut et al. (1996) exposed the appeal of internet content of local interest to what they identified as 'ordinary people'. More specifically, their project, which was a field trial based on providing internet access to people who were not familiar with the internet in mid 1990s, revealed that local online content attracted individuals online. Interestingly, despite the extensive effort to understand what drives individuals online, the attraction role of online content and, especially, of online content of local interest among other geographic, demographic and socio-economic factors has not been yet adequately explored. This paper is addressing this gap by utilising a novel source of web data, which is introduced in the next section.

3 Data

3.1 Local Internet Content

In order to measure the volume of Local Internet Content (LIC), we employ a novel source of archived webpages, which has never been used before in

such a context and extent. Specifically, we are utilising the JISC UK Web Domain Dataset, which is a subset of the Internet Archive, is curated by the British Library and includes all the archived webpages under the .uk top level domain[1]. Simply put, this is a list of billions of internet addresses (Uniform Resource Locator – URL) of .uk webpages, which have been archived during the period 1996-2013, as well as their archival timestamp (JISC and the Internet Archive, 2013). The contents of these webpages can be retrieved programatically via the Internet Archive[2]. The British Library has scanned the text of these archived webpages and created a separate subset of all the archived .uk webpages, which include a string in the format of a UK postcode (e.g. EC4A 2AH). This subset, which is used here to measure the volume of the LIC, includes 2.5 billion URLs and is known as the Geoindex data (Jackson, 2013).

The Internet Archive is a non-profit organisation, which aims to preserve digital content which otherwise would have been lost. It has been archiving web data since 1996 using a web crawler, which starts by archiving an initial list of URLs (seed list). During the archival process the hyperlinks to other URLs are also archived together with the content of the original URL. These hyperlinks are then used to find new URLs to archive, following a snowball-like process (Hale et al., 2017). Webpages can be archived multiple times over a year and more popular websites have higher probability of being archived and being archived more frequently (Hale et al., 2017). The Internet Archive only archives publicly available webpages and is also restricted to potential robot exclusions[3]. In 2016 the Internet Archive contained 273 billion webpages from 361 million websites, which took up 15 petabytes of storage (Internet Archive, 2016).

The recent literature includes some examples of business and innovation studies, which utilised web data from the Internet Archive. For instance, Papagiannidis et al. (2015) used such data to analyse the diffusion of web technologies and Papagiannidis et al. (2017) to build industrial classifications. Also, Blazquez and Domenech (2018) used archived web data from corporate websites to test the export orientation of a sample of Spanish companies, Arora et al. (2013) and Shapira et al. (2016) studied the early commercialization strategies of novel graphene technologies, Gök et al. (2015) explored the R&D activities and Li et al. (2016) created Triple Helix measures of green goods for small and mid-size enterprises. Musso and Merletti (2016) used the

[1] `http://data.webarchive.org.uk/opendata/ukwa.ds.2/`

[2] An example of an archived webpage retreived through the Internet Archive's web interface, known as the Wayback Machine, can be found in Appendix A

[3] These are standard exclusions policies used by websites to interact with other websites and web crawlers and are included in a robots.txt file

JISC UK Web Domain Dataset to rebuild the UK business web space for the period 1996-2001 and Hale et al. (2014) to analyse the British universities websites. Most of the above studies were limited in their scope (i.e. focused on small samples of archived web data) and all of them ignored the spatial signatures of these data. Moreover, no geographical studies, at least to our knowledge, have utilised these data.

Using the geoindex data we are able to build a measure of the volume of LIC. In a first step we preprocess the data in the following way: we filter the archived webpages to only include webpages with a valid British postcode. Then, we aggregate the observed webpages and their postcodes at the third level domain name and, therefore, our LIC measure is not based on the webpages with a UK postcode, but instead on the websites that these webpages belong to. For instance, if both `www.example.co.uk/page1` and `www.example.co.uk/page2` include one UK postcode, then we assign the two postcodes to the `example.co.uk` website. In other words, the postcode measurement takes place at the third level domain name[4]. Lastly, if we observe gaps in the data, we impute the postcodes for the missing years. Let the set of observed postcodes at year n for website `foo.co.uk` be p_n and the set of postcodes at year $n + x$ be p_{n+x}. Now if $p_n \subseteq p_{n+x}$, we use p for the years in-between n and $n + x$ as well.

After we preprocess the data, we calculate the volume of LIC. Recall that we are interested in assessing local online opportunities and not websites with a regional or national reach. We thus need a way to discount websites that have less of a local focus. To compute the geographic dispersion of a websites' set of postcodes p we calculate the Radius of Gyration r_g of p in kilometers. The Radius of Gyration is defined as follows (González et al., 2008):

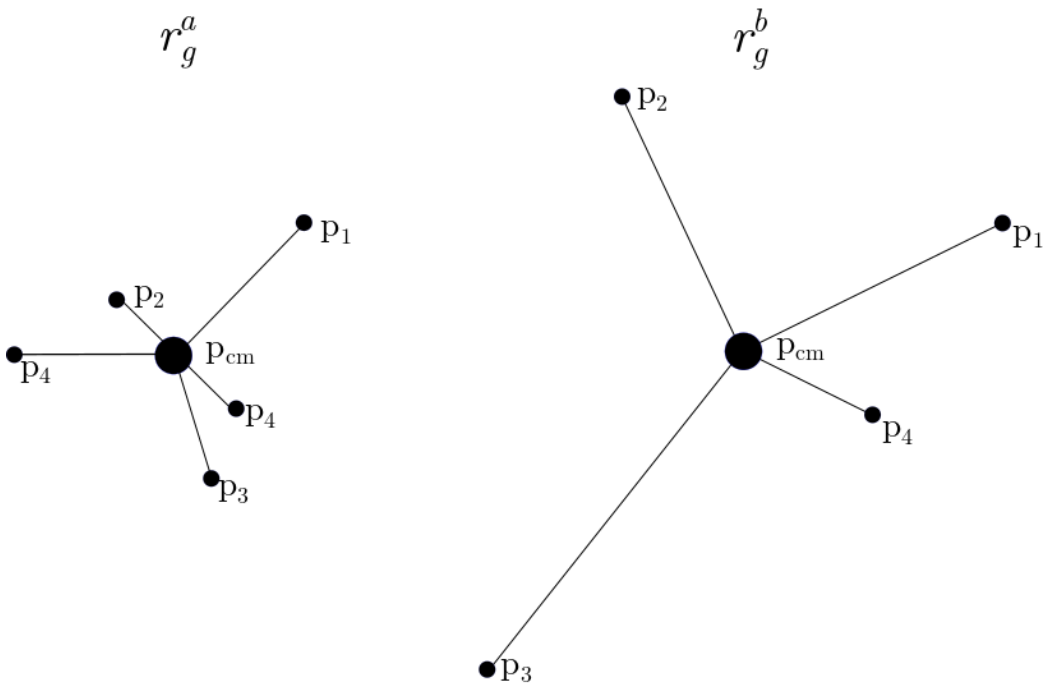
$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - p_{cm})^2},$$

where p_i represents the $i = 1, ..., n$ spatial coordinates of each postcode recorded for each domain and $p_{cm} = 1/n \sum_{i=1}^{n} p_i$ is the geographic center of mass of said domain. For an example of two r_g see figure 1.

A website with a high r_g will be of national interest, while a website with a low r_g will have a very local geographic presence. As local geographical units we utilise the Middle Layer Super Output Areas (MSOA) for England

[4]If .uk is the top level domain name and .co.uk is the second level domain name, then example.co.uk is the third level domain name.

Figure 1: An Example of Two Different r_g



$r_g^a < r_g^b$ as the average squared distance from the center of mass is much smaller for r_g^a than for r_g^b.

and the Intermediate Zones (IZ) for Scotland[5]. These are statistical units with a mean population of $7,200$.

For each MSOA/IZ with a set W of archived websites we calculate yearly measures of the volume of LIC as follows:

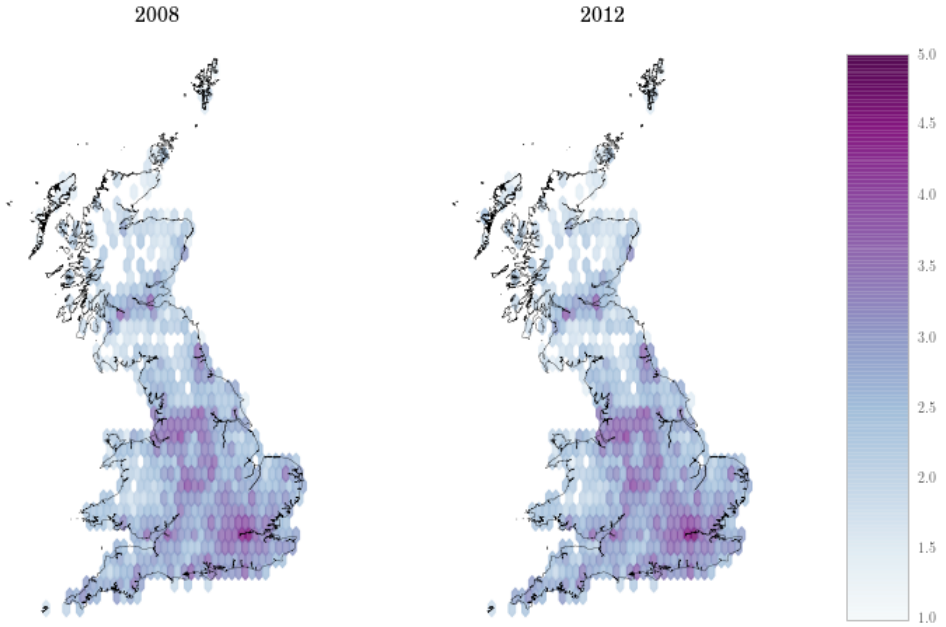
$$\sum_{p \in W} \frac{1}{1 + r_g(p)}$$

Figure 2) illustrates the volume of the $log_{10}(LIC)$ based on the above formula for the years 2008 and 2012. Because the distribution of this variable across MSOA/IZ is highly skewed we choose to plot the logarithm of the LIC. The spatial unit of the map is regular hexagons with the same centroids as the MSOA/IZ polygons. As the map illustrates London and the south-east region is undoubtedly the area with the highest volume of LIC. In addition, other large urban areas are also visible in this map, for example the Manchester-Liverpool-Leeds corridor as well as Birmingham. In total, the spatial distribution of the LIC reflects the British urban system. Note, that there are no dramatic changes between 2008 and 2012 regarding the volume of LIC, even though the underlying crawled websites differ considerably for the two years.

There is quite some debate in the literature regarding how much web content escapes the Internet Archive and, therefore, how much useful these data are for social science research. The critique has to do with the extent, the depth and the frequency of the archival process. In general, more popular webpages or, in other words, webpages with a lot of backlinks (i.e. hyperlinks landing on these webpages), have higher likelihood of being archived and, therefore, they are archived more often than less popular webpages. Nevertheless, the Internet Archive is the most extended archive in the world (Holzmann et al., 2016; Ainsworth et al., 2011). For example, Thelwall and Vaughan (2004) indicated that the Internet Archive included at least one webpage for 92 per cent of all the US commercial websites. Moreover, Hale et al. (2017) assessed the 'depth' of the Internet Archive by comparing the 'live' and archived Trip Advisor London webpages. Although it is not easy to generalise the results based on a single website, they found that only 24 per cent of these webpages had been archived. Again, webpage popularity appeared to drive the archival bias. Earlier work from Ainsworth et al. (2011) indicated that between 35 and 90 per cent of webpages have been archived by public archives including, but not limited to, the Internet Archive. Moreover, estimations indicated that around 70 per cent of all websites contain place reference (Hill, 2009).

[5]Northern Ireland was excluded from the analysis as the available census geographies were not comparable in population.

Figure 2: Local Internet Content in Great Britain (log_{10})



Hence, our data might be biased towards more popular websites. We partially address this issue with our weighting process as the radius of gyration assigns higher weight to websites of local interest. Therefore, websites such as Trip Advisor with webpages including a lot of postcodes around the country will be assigned a very low weight. Nevertheless, websites of local internet which are anchored to larger cities and serve larger populations can be more popular in absolute terms than websites which serve smaller cities. Because of the above, we expect that the LIC measure will capture a lower share of all the websites of local interest in less populated areas and, therefore, a downward bias of the LIC regression coefficient for such areas.

Regarding the temporal frequency of the archival process, one might argue that a webpage is not archived often enough to have yearly measures of LIC. However, this does not seem to be a problem. For the $5,740,059$ archived websites which had a webpage with a UK postcode during the period 1996-2013, only $432,291$ (i.e. 7.5 per cent) appeared in more than one years without these two years being consequent, and the imputation process described above addresses this issue.

3.2 British Population Survey

The second main source of data we employ in order to expose the drivers of individual online behaviour is the British Population Survey (BPS). The BPS is a monthly rolling market research survey. It covers the period 2008 to 2015 and each month about 6000 to 8000 individuals are surveyed, thus leading to over 80,000 participants per year. The sampling methodology provides an accurate cross-section of the British population at the postcode level. While half of the sample is based on geodemographic models, the other half focuses on under-weighted profiles to increase the representativeness of the survey[6]. As our last fully observed year for the LIC is 2012, we only use the data from the years 2008 to 2012 inclusive for our analysis. Data is collected on a variety of socio-economic topics such as demographics, economics, shopping preferences, durables, and media and internet usage. Most importantly, the BPS not only collects data about the frequency of internet usage as well as other individual characteristics, but also the postcode of all respondents, something which enables us to match individuals with their LIC[7].

4 Methods

In order to reveal the individual and neighbouring characteristics that affect online behaviour, we employ multilevel models (MLM), which have the capacity to control for the clustering of individual observations within the same contextual (i.e. geographic) unit. This attribute reflects economic reality as individuals are hardly ever isolated from their direct environment, but they are nested within complex economic and institutional contexts (Hundt and Sternberg, 2016). MLM have been widely used in human geography. For instance, Loo et al. (2017) employed MLM in order to explore how the environment at the neighbourhood level is related to physical and mental health of senior population. In a different study, Bakke et al. (2009) used MLM in order to understand the individual and contextual factors that affect forgiveness in conflict-affected societies. Similarly, López-Bazo and Motellón (2018) investigated the role that firm and regional level characteristics play in innovation performance of firms using MLM.

Our MLM contain two levels of variables. Level 1 variables are tied to an individual and vary at the individual level. Such variables include, for example, individual income and gender, and are part of the BPS. Level 2 variables

[6]The content as well as the procedure for accessing the BPS are described in detail at `https://data.cdrc.ac.uk/dataset/british-population-survey`.

[7]Because of their sensitive nature, the matching of the survey with the other data and the analysis described in the next section took place in a secure server environment.

capture the effects of the geographical context, within which individuals are embedded, on individual online behaviour. To capture this geographical context we adopt MSOA/IZ as Level 2 of our modelling exercise. Therefore, we employ variables at this level such as population density and average house prices. Moreover, this is the level that we aggregate the LIC measure and match it with individual observations.

In general, there are three options for designing MLM: (i) random intercepts models in which the intercepts are allowed to vary between groups, (ii) random slopes models, wherein the slopes are allowed to vary between groups, and (iii) random intercepts and slopes models, where both intercepts and slopes are allowed to vary between groups (Snidjers and Bosker, 2003). While a random intercepts and slopes model might be the most comprehensive type of model, it is also the most complex one as it has the most coefficients.

Our focus on a local phenomenon – LIC – leads to a relatively large number of areas with a relatively low amount of observations (over half of the observed areas have 28 or fewer observations for the whole study period). Given the sparsity of our data and the failure of a model that uses both random intercepts and random slopes to converge, a more realistic option was to either use a random intercepts or random slopes model. As different geographic areas have varying levels of internet usage to begin with, we have decided to use a random intercepts model. Hence, our random intercept model for testing our hypothesis for the individual i in the area j is defined as follows:

$$\begin{aligned} Internet\ usage_{ij} = \gamma_{00} + \gamma_{10} Individual_{ij} + \gamma_{01} LIC_j + \gamma_{02} Area_j + \\ \gamma_{11} Year + e_{0j} + e_{ij} \end{aligned} \tag{1}$$

where *Internet usage* refers to our dependent variable, *Individual* to all independent variables at the level of the individual, *LIC* to our metric of local internet content, and *Area* to all independent variables that pertain to the area an individual is located in (see also Table 2 for more details). Note, that all of the independent variables are z-score standardized.

Our dependent variable is the *Internet usage* of individual i located in MSOA/IZ j. As frequency of internet usage was originally encoded in an ordinal scale in the BPS, we cannot simply use the frequency of internet usage as our dependent variable for a regression model. Unfortunately given the high class imbalance of the orignal ordinal variable and the relatively high amount of areal units it is not feasible to estimate an ordinal model. Logistic regression and by extension ordinal regression models significantly misjudge the probability of rare events (King and Zeng, 2001). We have thus

opted to construct a binomial dependent variable of frequent internet usage vs. non-frequent internet usage to be able to run logistic regression models.

Our main aim is to estimate the effect of the LIC available to an individual i located in MSOA/IZ j on their *Internet usage*. On top of this, we also control for a variety of socio-demographic individual and geographical characteristics that we know from the literature (see Section 2) that can affect individual internet usage. At the individual level we thus use gender, household income, age, and qualification as confounding variables, where both household income and qualifications are ordinal variables (from low to high). Table 1 provides a description and Table 2 presents the descriptive statistics of all the variables included in the model. To account for the more intense internet usage and activity in cities, we also include in the regressions the distance between the MSOA/IZ of each individual and the center of the closest urban area as defined by the ONS[8]. As London is arguably the most important urban centre and the only global city in Great Britain we also include the distance to London as another control variable. To further account for the spatial structure and characteristics, we include in our models population density at the MSOA/IZ level. Furthermore, as more affluent areas are associated with more frequent internet usage, we also include the average house price in an area as an independent variable[9]. To account for spatial dependency in our data, we also include the spatial lags of house prices, population density, and the LIC as independent variables. For every MSOA/IZ area j we calculate the spatial lag by averaging all areas that are adjacent to j (Queen contiguity) for population density, house prices and LIC. Spatial lags allow to capture not only the immediate neighbourhood of an individual, but also the wider geographic setting within which they are embedded. For example, while area j might densely populated, the surrounding areas determine whether it is located in a more rural or urban setting. In addition, following previous research we also control for the quality of the internet infrastructure at the MSOA/IZ level. To do so, we include a variable for the distance between the MSOA/IZ an individual resides to the nearest internet exchange[10]. As we know from the literature, internet speed drops with increasing distance to exchanges (e.g. Riddlesden and Singleton, 2014; Nardotto et al., 2015). Last but not least, we also account for the fact that internet usage becomes more frequent with time by including a time trend variable $Year$.

[8]http://www.nomisweb.co.uk/articles/ref/builtupareas_userguidance.pdf

[9]House prices and population density were sourced from `www.ons.gov.uk` and `statistics.gov.scot` for England and Wales, and Scotland respectively.

[10]For the location of internet exchanges we are using the data from (Nardotto et al., 2015)

Table 1: Description Variables

Variable	Description	Domain
	Individual level variables	
Internet usage	Frequency of internet usage	{frequent user, not frequent user}
Gender	Gender	{male, female}
Income	Household income in GBP	{[0, 4499], [4500, 6499], [6500, 7499], [7500, 9499], [9500,11499], [11500, 13499], [13500, 15499], [15500, 17499], [17500, 24999], [25000, 29999], [30000, 39999], [40000, 49999], [50000, 74999], [75000, 99999], [100000, ∞)}
Age	Numeric age	$[0, \infty)$
Qualifiaction	Level of education	{A-level, Bachelor, GCSE/O-LEVEL/CSE, Masters/PhD, no formal qualifiactions, still studying, vocational qualifiactions}
	Area level variables	
Population density	Inhabitants per km^2 per MSOA	$[0, \infty)$
House prices	The aveage house price per MSOA	$[0, \infty)$
LIC	Local internet content	$[0, \infty)$
Distance to urban center	Spatial distance to the closest conurbation	$[0, \infty)$
Distance to London	Spatial distance to London	$[0, \infty)$
Distance to exchange	Spatial distance to the closest Internet exchange	$[0, \infty)$
	Yearly trend	
Year	Year the individual was questioned	[2008, 2012]

Table 2: Descriptive statistics

	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Individual level variables							
Internet usage	397,715	198,857.500	58,823.506	0	0	1	1
Gender (Male)	397,715	198,857.500	2044.2457	0	0	1	1
Income	247,477	7.128	3.877	1.000	4.000	10.000	15.000
Age	397,715	47.656	19.677	0	31	64	99
Qualification	362,398	3.482	1.855	1.000	2.000	5.000	7.000
Area level variables							
Population density	397,715	3,408.968	3,350.000	2.118	848.955	4,673.784	26,284.970
SL(Pop. dens.)	397,715	2,997.266	2,731.171	2.760	1,029.646	4,182.609	19,147.650
House prices	397,715	168,069.800	74,313.000	18,000	115,000	208,500	1,400,000
SL(House prices)	397,690	175,931.000	67,191.570	54,900.000	123,939.600	215,800.000	1,165,500.000
LIC	397,715	59.331	58.139	0.019	23.418	79.114	1,598.645
SL(LIC)	397,715	63.326	44.798	3.890	37.336	79.640	1,827.741
Dist. to urban center	397,715	17,081.350	15,784.350	18.639	7,093.552	21,727.830	278,891.300
Distance to London	397,715	210,273.900	156,650.000	1,280.526	83,108.000	281,047.900	861,116.800
Distance to exchange	397,715	1,715.158	1,224.333	15.175	948.568	2,189.614	10,729.980

5 Results

Table 3 presents the outputs of the MLM. Each column includes a single regression, the dependent variable of which is always a binomial dependent variable representing frequent internet usage (1) versus non-frequent usage (0). Column 1 only includes the intercept and the corresponding random effect for the MSOA/IZ, which shows that our data exhibit significant clustering at our chosen level of geography. Specifically, the Intraclass Correlation Coefficient (ICC) indicates that 14.4 per cent of the individual variability in internet usage can be accounted to the location of individuals in MSOA/IZ. Then, we incrementally include in the model individual level fixed effects, geographical fixed effects, geographical fixed effects including also spatial lags and, finally, an all-inclusive model with all the individual and geographical characteristics.

The model with only the individual effects is presented in Column 2 and most of them are in line with previous findings from the relevant literature. For instance, higher income increases the likelihood of an individual being a frequent internet user, while the opposite seems to happen with age. Previous findings from empirical studies have identified the positive and negative role, respectively, of income and age on internet usage. Also, qualifications seem to have a negative effect on the frequency of internet usage, which may be attributed to the disconnect between frequent internet usage and capital-enhancing activities. Moreover, the effect of the time trend variable is positive, something which illustrates and controls for the fact that the likelihood of an individual being a frequent internet user increases over time. Lastly, our results indicate that the likelihood of being a frequent internet user is higher for male than female respondents. Although previous studies illustrated a more nuanced picture of gender-related typologies of internet usage, our results support previous findings that men are more active online. This effect is significant and consistent throughout the different specifications discussed here.

Columns 3 and 4 include the geographical effects and the spatial lags for some of them. The main effect of interest is the LIC. As our model outputs indicate, the likelihood of an individual being a frequent internet user is positively associated with the availability of online content of local interest. Our interpretation is that online content of local interest represents online opportunities for individuals and, therefore, the availability of such content attracts individuals to spend more time online.

This headline result is consistent against different specifications and remains stable after controlling for various other variables. Firstly, LIC remains a significant predictor of the frequency of internet usage even after the in-

Table 3: Results MLM Models

	Dependent variable:				
	Internet usage				
	(1)	(2)	(3)	(4)	(5)
Individual level variables					
Gender (Male)		0.390***			0.390***
Income		0.053***			0.050***
Age		−0.846***			−0.851***
Qualitifcation		−0.333***			−0.326***
Area level variables					
Population density			0.147***	0.028	−0.063***
SL(Pop. dens.)				0.131***	0.001
House prices			0.096***	0.107***	0.257***
SL(House prices)				−0.053***	−0.004
LIC			0.120***	0.073***	0.028**
SL(LIC)				0.095***	−0.001
Dist. to urban center			−0.034***	−0.020*	−0.047***
Distance to London			−0.076***	−0.075***	−0.041***
Distance to exchange			0.004	−0.006	−0.032**
Yearly trend					
Year		0.328***			0.333***
Constant					
Constant	−0.460***	−0.559***	−0.450***	−0.450***	−0.582***
ICC	0.144	0.149	0.124	0.127	0.122
Observations	397,715	228,176	397,715	397,690	228,176
Log Likelihood	-248,718.100	-128,866.900	-248,298.800	-248,216.700	-128,496.700
Akaike Inf. Crit.	497,440.200	257,747.900	496,613.700	496,455.400	257,025.400
Bayesian Inf. Crit.	497,462.000	257,820.300	496,700.800	496,575.200	257,190.800

Notes: *p<0.1; **p<0.05; ***p<0.01
Spatial lags (SL) are the average of neighboring MSOAs
Ind. var. are z-score standardized

clusion of its spatial lag in the model. As Column 4 illustrates both the LIC available in the MSOA/IZ where an individual is located, but also the average LIC in the surrounding MSOA/IZ have a positive and significant effect on the likelihood of that individual being a frequent internet user. Interestingly, the coefficient of the spatial lag of the LIC variable is slightly larger than the one of the actual LIC variable. This highlights, for example, that the availability of LIC in a broader urban area would have a larger effect on the likelihood of an individual being a frequent internet user than the volume of LIC in their direct neighbourhood. Nevertheless, when both individual and geographical variables are included in the regressions (Column 5) only the coefficient of the actual LIC remains significant. It needs to be highlighted here that due to missing data for some of the individual level variables the regressions presented in Columns 4 and 5 include more observations and, therefore, a different sample than the rest of the regressions. The latter might explain the lack of a significant coefficient for the spatial lag of the LIC. Nevertheless, even when the regression includes less observations as in Column 5 and also individual level variables, the LIC remains a significant predictor of the individual internet usage.

Importantly, the effect of LIC is consistent against the inclusion of other control variables. Columns 3 and 4 of Table 3 also include other geographical control variables that we expect to affect individual internet usage. Indeed, population density appears to be positively related with individual internet usage. Similarly to LIC, when both population density and its spatial lags are included in the regression, only the latter appears to have a significant positive effect. However, this effect is not consistent. Column 5, which includes both the individual and geographical effects and, therefore, is our preferred specification, illustrates a negative effect of population density on the likelihood of frequent internet usage. Although this finding might sound contradicting, in reality reflects the attributes of the UK spatial structure, where high income population groups can be found in low density non-urban places. Hence, after controlling for individual income, which has a consistent positive effect, urban density appears to be negatively related with frequent internet usage.

The above argument is also reflected in the two variables controlling for the location of the MSOA/IZ. Indeed, we are including as explanatory variables the distance of each MSOA/IZ to London and to the nearest urban centre. By using these two variables we are able to control for how proximity to an urban centre is related to individual online behaviour. Given also the unique role that London performs within the UK urban system, we are testing whether there is a London-specific effect. As expected, both of these variables are negatively related with the frequency of internet usage. The fur-

ther away an MSOA/IZ is from London or from an urban centre, the lower is the likelihood of an individual residing there to be a frequent internet user. This is in line with early findings from internet geography studies, which highlighted the urban character of internet – both in terms of infrastructure and usage – especially at the early stages of its commercialisation.

We are also controlling for the quality of the internet infrastructure by including a variable, which measures the distance of the MSOA/IZ an individual resides to the nearest internet exchange. Again, our findings are in agreement with the ones from previous studies presented in Section 2. The further away an MSOA/IZ is located from an internet exchange and, therefore, the lower the expected internet speed is for users located in this MSOA/IZ, the lower is the likelihood of an individual residing there to be a frequent internet user.

Lastly, we are also controlling for the average house price in the MSOA/IZ of an individual as well as the the average house price in the adjacent MSOA/IZ (spatial lag). While there is a consistent and positive effect of the former, this is not the case for the spatially lagged variable. However, the results of our preferred specification in Column 5 indicate that the frequency of internet usage is positively correlated with house prices. These variables represent the overall affluence of the area an individual resides. Interestingly, such a neighbourhood effect remains statistically significant even after the inclusion of individual income. This highlights the important role place performs in affecting the online behaviour of individuals.

Although a causal interpretation of the above results is beyond the scope of this paper, effort is spent in order to address two potential endogeneity issue. First, one might argue that the variables in the final model as depicted in Column 5 do not account for all spatial variation in the data as the ICC is still comparatively large for the final model. To investigate whether our final model adequately captures the spatial variation in internet usage, we test whether the residuals of the model are spatially auto-correlated. While we find significant spatial auto-correlation for the residuals, the observed Moran's I statistic is only ~ 0.05. Indicating at best a very weak positive spatial auto-correlation for the residuals and showing that our model adequately captures the spatial variation in our data.

Second, one can claim that the above models are affected by reverse causality as the increased frequency of internet usage can lead to the production of more online content of local interest. Such sizeable feedback effects could have been expected for social media and other types of user generated online content. Nevertheless, we know from related studies that only a very small share of internet users are heavy content creators, while the vast majority of internet users are content consumers (Graham and Shelton,

2013; Haklay, 2013; Graham, 2014). In addition, the JISC UK Web Domain data we used for the LIC variable only include archived webpages under the .uk domain name excluding most of the widely used social media platforms, which are hosted under the .com domain name (e.g. facebook.com, twitter.com, linked.com). Hence, we expect that is less likely for the above models to suffer from such a reverse causality. In order to further address this potential issue we are estimating again the preferred specification, but instead of using the contemporaneous LIC, we are utilising a one year lag of LIC and its spatial lag. The results, which are presented in Table 4, are qualitatively similar to the ones presented in Column 5 from Table 3. Interestingly, the coefficient of the lagged LIC variable is marginally larger than the one for the contemporaneous LIC reflecting a higher correlation between the frequency of internet usage and the past availability of online content of local interest. The specification presented in Table 4 can partially address such a reverse causality issue as the lagged volume of LIC cannot be affected by future individual online behaviour.

In a nutshell, the above results indicate a statistically significant and consistent positive relationship between the volume of online content of local interest and the frequency of individual internet usage. Importantly, our results are robust against different specifications and remain significant even after controlling for various factors, the importance of which has been highlighted by previous studies.

6 Conclusion and Discussion

The aim of this paper is to test whether the availability of web content of local interest can attract individuals online. Our underpinning assumption is that local internet content (LIC) in the form of websites with specific geographical reference represents online opportunities relevant at the local scale and, therefore, it can act as a pull factor for individuals to spend more time online. Such online opportunities can involve, for example, consumption opportunities of services and products provided by local businesses, governmental services such as services by local authorities, or third sector organisations present in a locality. Although we know a lot about the individual (e.g. income) and geographical (e.g. population density) factors that push individuals online as well as the non-geographical online opportunities that can pull individuals online (e.g. large online retailers or central government services), the literature has not yet explored whether online content of local interest also attracts individuals online.

In order to answer the above research question we employed novel data

Table 4: Results Lag Model

	Dependent variable:
	Internet usage
Individual level variables	
Gender (Male)	0.390^{***}
Income	0.050^{***}
Age	-0.851^{***}
Qualitifcation	-0.326^{***}
Area level variables	
Population density	-0.059^{**}
SL(Pop. dens.)	$-0.002.$
House prices	0.257^{***}
SL(House prices)	-0.003
TL(LIC)	0.037^{***}
TL(SL(LIC))	$-0.011.$
Dist. to urban center	-0.047^{***}
Distance to London	-0.041^{***}
Distance to exchange	-0.031^{**}
Yearly trend	
Year	0.332^{***}
Constant	
Constant	-0.582^{***}
ICC	0.122
Observations	228,176
Log Likelihood	-128,494.400
Akaike Inf. Crit.	257,020.700
Bayesian Inf. Crit.	257,186.100
Note:	*p<0.1; **p<0.05; ***p<0.01
	Temporal lags (TL) are the value of the previous year
	Spatial lags (SL) are the avg. of neighboring MSOAs

from the Internet Archive, which contain all the archived webpages under the .uk top level domain. Using the text from the webpages we are able to geolocate these webpages and only include in the analysis the ones which contain a valid UK postcode. We process these data in way that enables us to decrease the weight of websites that have less of a local focus. Then, we calculate the volume of online content of local interest at the MSOA/IZ level and spatially match it with a large population survey, which includes information about the individual online usage as well as other individual characteristics and also their postcode. Therefore, we are able to test whether the availability of LIC in proximity to the individual location is related to the frequency of internet usage. Importantly, we are able to do so by controlling also for other individual and geographical characteristics that previous research has identified as strong predictors of internet usage.

Our statistical analysis revealed that the volume of LIC is a significant predictor of the frequency of internet usage. This result is consistent against different specifications and is not sensitive to the inclusion of different control variables regarding both individual and geographical characteristics. This is the first time, at least to our knowledge, that we are able to connect internet usage with the availability of online content at the local level.

The exposure of this relationship can provide new insights on digital divides. As previous research indicated, digital divides tend to be clustered in space (Hindman, 2000; Graham, 2011). Such divides can impede any objectives for social inclusion and economic efficiency given how much our society is structured around the internet (Sparks, 2013). Importantly, the literature has moved beyond the deterministic idea that internet connectivity can automatically alleviate any divides regarding the benefits that derived from the internet. Although physical access to the internet can still be an important issue, it is not necessarily the primary driving force behind digital divides, at least for Western countries (Scheerder et al., 2017). For instance, internet access is not a problem for 90 per cent of the UK households in 2018 (ONS, 2018). Therefore, there is a call for understanding digital divides not merely as physical access to the internet, but also in terms of the necessary skills individuals have in order to access the internet and, most recently, in terms of the benefits linked to internet usage (Fuchs, 2009; Selwyn, 2004; Van Dijk, 2005; Scheerder et al., 2017). Wei et al. (2011) termed this last dimension *the third level digital divide*. This paper proposes the idea that the benefits an individual enjoys by using the internet are related, among other things, to the availability of online content of local interest. Our analysis indicated that internet content of local interest represents a pool opportunities and, therefore, individuals spend more time online in order to access these opportunities. Hence, our efforts to tackle digital divides should not be

monopolised by internet connectivity related measures and subsidies. On the contrary digital divide policies should also be accompanied by a demand-side dimension in order to support the creation of online content from local actors (e.g. businesses), something which can attract more individuals online.

Acknowledgements

We want to acknowledge the following data providers: (i) The British Population Survey data have been produced by DataTalk and made available to us by the Consumer Data Research Centre (CDRC), an ESRC Data Investment (Project Numbers: ES/L011840/1 and ES/L011891/1). (ii) The JISC UK Web Domain Dataset has been developed by the British Library and the UK Web Archive. (iii) The locations of internet exchanges have been kindly provided by Dr Mattia Nardotto (see Nardotto et al., 2015). The research for this paper has been funded by a CDRC Innovation Fund grant and a University of Birmingham ESRC Impact Accelerator Account grant.

References

Agarwal, R., Animesh, A., and Prasad, K. (2009). Research note—social interactions and the "digital divide": Explaining variations in internet use. *Information Systems Research*, 20(2):277–294.

Ainsworth, S. G., Alsum, A., SalahEldeen, H., Weigle, M. C., and Nelson, M. L. (2011). How much of the web is archived? In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 133–136. ACM.

Arora, S. K., Youtie, J., Shapira, P., Gao, L., and Ma, T. (2013). Entry strategies in an emerging technology: a pilot web-based study of graphene firms. *Scientometrics*, 95(3):1189–1207.

Bakke, K. M., O'Loughlin, J., and Ward, M. D. (2009). Reconciliation in conflict-affected societies: Multilevel modeling of individual and contextual factors in the north caucasus of russia. *Annals of the Association of American Geographers*, 99(5):1012–1021.

Bekkerman, A. and Gilpin, G. (2013). High-speed Internet growth and the demand for locally accessible information content. *Journal of Urban Economics*, 77:1–10.

Blank, G., Graham, M., and Calvino, C. (2018). Local geographies of digital inequality. *Social Science Computer Review*, 36(1):82–102.

Blank, G. and Groselj, D. (2014). Dimensions of internet use: amount, variety, and types. *Information, Communication & Society*, 17(4):417–435.

Blazquez, D. and Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy*, 24(2):406–428.

Bughin, J., Hazan, E., Manyika, J., and Woetzel, J. (2016). Digital Europe: Pushing the Frontier. *Capturing the Benefits", McKinsey Global Institute*.

Calenda, D. and Meijer, A. (2009). Young people, the internet and political participation: findings of a web survey in italy, spain and the netherlands. *Information, Communication & Society*, 12(6):879–898.

Di Gennaro, C. and Dutton, W. (2006). The internet and the public: Online and offline political participation in the united kingdom. *Parliamentary affairs*, 59(2):299–313.

Farag, S., Schwanen, T., Dijst, M., and Faber, J. (2007). Shopping online and/or in-store? A structural equation model of the relationships between e-shopping and in-store shopping. *Transportation Research Part A: Policy and Practice2*, 41:125–141.

Forman, C., Goldfarb, A., and Greenstein, S. (2018). How geography shapes—and is shaped by—the internet. *The New Oxford Handbook of Economic Geography*, page 269.

Fuchs, C. (2009). The role of income inequality in a multivariate cross-national analysis of the digital divide. *Social Science Computer Review*, 27(1):41–58.

Fuchs, C., Hofkirchner, W., Schafranek, M., Raffl, C., Sandoval, M., and Bichler, R. (2010). Theoretical foundations of the web: cognition, communication, and co-operation. Towards an understanding of Web 1.0, 2.0, 3.0. *Future Internet*, 2(1):41–59.

Gök, A., Waterworth, A., and Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, 102(1):653–671.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns supplementary material. *Nature*, 453.

Gottmann, J. (1977). Megalopolis and antipolis: The telephone and the structure of the city. In Pool, I. d. S., editor, *The social impact of the telephone*, pages 303–317. MIT Press, Cambridge, MA.

Graham, M. (2011). Time machines and virtual portals: The spatialities of the digital divide. *Progress in development studies*, 11(3):211–227.

Graham, M. (2014). Inequitable distributions in internet geographies: The global south is gaining access, but lags in local content. *Innovations: Technology, Governance, Globalization*, 9(3-4):3–19.

Graham, M., De Sabbata, S., and Zook, M. A. (2015a). Towards a study of information geographies:(im) mutable augmentations and a mapping of the geographies of information. *Geo: Geography and environment*, 2(1):88–105.

Graham, M. and Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, 3(3):255–261.

Graham, M., Straumann, R. K., and Hogan, B. (2015b). Digital divisions of labor and informational magnetism: Mapping participation in wikipedia. *Annals of the Association of American Geographers*, 105(6):1158–1178.

Grubesic, T. H. (2008). The spatial distribution of broadband providers in the United States: 1999–2004. *Telecommunications Policy*, 32(3):212–233.

Haklay, M. (2013). Neogeography and the delusion of democratisation. *Environment and Planning A*, 45(1):55–69.

Hale, S. A., Blank, G., and Alexander, V. D. (2017). Live versus archive: Comparing a web archive to a population of web pages. In Brügger, N. and Schroeder, R., editors, *Web as History: Using Web Archives to Understand the Past and the Present*, pages 45–61. UCL Press, London.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., and Margetts, H. (2014). Mapping the UK webspace: Fifteen years of british universities on the web. In *Proceedings of the 2014 ACM conference on Web science*, pages 62–70. ACM.

Hill, L. L. (2009). *Georeferencing: The geographic associations of information*. Mit Press.

Hindman, D. B. (2000). The rural-urban digital divide. *Journalism & Mass Communication Quarterly*, 77(3):549–560.

Hoffmann, C. P., Lutz, C., and Meckel, M. (2015). Content creation on the internet: A social cognitive perspective on the participation divide. *Information, Communication & Society*, 18(6):696–716.

Holzmann, H., Nejdl, W., and Anand, A. (2016). The Dawn of today's popular domains: A study of the archived German Web over 18 years. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference*, pages 73–82. IEEE.

Hundt, C. and Sternberg, R. (2016). Explaining new firm creation in Europe from a spatial and time perspective: A multilevel analysis based upon data of individuals, regions and countries. *Papers in Regional Science*, 95(2):223–257.

Internet Archive (2016). Internet archive blogs. `https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/`. Accessed: 2010-09-30.

Jackson, A. N. (2013). Jisc uk web domain dataset (1996-2010) geoindex.

JISC and the Internet Archive (2013). Jisc uk web domain dataset (1996-2013).

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9:137–163.

Kitchin, R. and Dodge, M. (2011). *Code/space: Software and everyday life*. Mit Press.

Kraut, R., Scherlis, W., Mukhopadhyay, T., Manning, J., and Kiesler, S. (1996). Homenet: A field trial of residential internet services. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 284–291. ACM.

Li, Y., Arora, S., Youtie, J., and Shapira, P. (2016). Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation*, In Press.

Longley, P. A. and Singleton, A. D. (2009). Classification through consultation: Public views of the geography of the e-society. *International Journal of Geographical Information Science*, 23(6):737–763.

Longley, P. A., Webber, R., and Li, C. (2008). The uk geography of the e-society: a national classification. *Environment and Planning A*, 40(2):362–382.

Loo, B. P., Lam, W. W., Mahendran, R., and Katagiri, K. (2017). How is the neighborhood environment related to the health of seniors living in hong kong, singapore, and tokyo? some insights for promoting aging in place. *Annals of the American Association of Geographers*, 107(4):812–828.

López-Bazo, E. and Motellón, E. (2018). Innovation, heterogeneous firms and the region: evidence from spain. *Regional Studies*, 52(5):673–687.

Mack, E. A. and Grubesic, T. H. (2009). Forecasting broadband provision. *Information Economics and Policy*, 21(4):297–311.

Malecki, E. J. (2002). The economic geography of the Internet's infrastructure. *Economic Geography*, 78(4):399–424.

Mills, B. F. and Whitacre, B. E. (2003). Understanding the non-metropolitan—metropolitan digital divide. *Growth and Change*, 34(2):219–243.

Musso, M. and Merletti, F. (2016). This is the future: A reconstruction of the UK business web space (1996–2001). *New media & society*, 18(7):1120–1142.

Nardotto, M., Valletti, T., and Verboven, F. (2015). Unbundling the incumbent: Evidence from uk broadband. *Journal of the European Economic Association*, 13(2):330–362.

ONS (2018). Internet access – households and individuals, great britain: 2018. `https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2018`. Accessed: 2018-11-26.

Oughton, E., Tyler, P., and Alderson, D. (2015). Who's Superconnected and Who's Not? Investment in the UK's Information and Communication Technologies (ICT) Infrastructure. *Infrastructure Complexity*, 2(1):1–17.

Papagiannidis, S., Gebka, B., Gertner, D., and Stahl, F. (2015). Diffusion of web technologies and practices: A longitudinal study. *Technological Forecasting and Social Change*, 96:308–321.

Papagiannidis, S., See-To, E. W. K., Assimakopoulos, D. G., and Yang, Y. (2017). Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the Internet age? *Computers & Operations Research*, In Press.

Riddlesden, D. (2014). Internet User Classication (IUC) User Guide.

Riddlesden, D. and Singleton, A. D. (2014). Broadband speed equity: A new digital divide? *Applied Geography*, 52:25–33.

Scheerder, A., van Deursen, A., and van Dijk, J. (2017). Determinants of internet skills, uses and outcomes. a systematic review of the second-and third-level digital divide. *Telematics and Informatics*.

Selwyn, N. (2004). Reconsidering political and popular understandings of the digital divide. *New media & society*, 6(3):341–362.

Shapira, P., Gök, A., and Salehi, F. (2016). Graphene enterprise: mapping innovation and business development in a strategic emerging technology. *Journal of Nanoparticle Research*, 18(9):269.

Sinai, T. and Waldfogel, J. (2004). Geography and the Internet: is the Internet a substitute or a complement for cities? *Journal of Urban Economics*, 56(1):1–24.

Singleton, A. D., Riddlesden, D., Blank, G., and Graham, M. (2015). Internet User Map Book.

Snidjers, T. A. B. and Bosker, R. J. (2003). Multilevel Analysis. An introduction to basic and advanced multilevel modeling.

Sparks, C. (2013). What is the "digital divide" and why is it important? *Javnost-The Public*, 20(2):27–46.

Thelwall, M. and Vaughan, L. (2004). A fair history of the web? examining country balance in the internet archive. *Library & information science research*, 26(2):162–176.

Tranos, E. (2013). *The geography of the internet: Cities, regions and internet infrastructure in Europe*.

Van Deursen, A. J. and Van Dijk, J. A. (2014). The digital divide shifts to differences in usage. *New media & society*, 16(3):507–526.

Van Dijk, J. A. (2005). *The deepening divide: Inequality in the information society*. Sage Publications.

Viard, V. B. and Economides, N. (2014). The effect of content on global internet adoption and the global "Digital divide". *Management science*, 61(3):665–687.

Wei, K.-K., Teo, H.-H., Chan, H. C., and Tan, B. C. (2011). Conceptualizing and testing a social cognitive model of the digital divide. *Information Systems Research*, 22(1):170–187.

Wheeler, D. C. and O'Kelly, M. E. (1999). Network topology and city accessibility of the commercial Internet. *Professional Geographer*, 51(3):327–339.

Zillien, N. and Hargittai, E. (2009). Digital distinction: Status-specific types of internet usage. *Social Science Quarterly*, 90(2):274–291.

A Appendix

Figure 3: Example of an archived version of Bloomberg.com from 2015